

BUPT-MCPRL at TRECVID 2011*

Zhicheng Zhao, Yanyun Zhao, Xin Guo, Yuanbo Chen, Yan Hua, Wen Wang, Cheng Liu, Siyuan Wu, Han Zhang, Lingxi Wang, Yuanhui Mao, Anni Cai, Menghua Zhai
Multimedia Communication and Pattern Recognition Labs,
Beijing University of Posts and Telecommunications, Beijing 100876, China
{zhaozc, zyy, annicai}@bupt.edu.cn

Abstract

In this paper, we describe BUPT-MCPRL systems for TRECVID 2011. Our team participated in five tasks: semantic indexing, known-item search, instance search content-based copy detection and surveillance event detection. A brief introduction is shown as follows:

A. Known-item search

In this year, we proposed two different methods: one based on text and another is bio-inspired method. All 2 runs we submitted are described in Table 1.

Table 1. KIS results and descriptions for each run

Run ID	Mean Inverted Rank	Description
F_A_YES_MCPRBUPT1_1	0.455	This run is based on text.
F_A_NO_MCPRBUPT2_2	0.009	This run is based on visual attention model and concept/object detection.

B. Instance search

This year, we mainly focused on the following parts: selection of distance metric, multimodal fusion and results re-ranking, and finally, we submitted 4 runs and achieved a high infAP.

C. Semantic indexing

In this task, different SIFT-like features were tested and 3 fusion strategies were adopted. However, the infMAP was dissatisfactory.

D. Content-based copy detection

Two approaches for the content-based copy detection task were proposed, and PIP was detected independently.

E. Surveillance event detection

This year, we mainly evaluated the events of PersonRuns, PeopleMeet, PeopleSplitUp, ObjectPut, Embrace and Pointing. Our system adopted different algorithms in detecting these events accordingly.

1 Known-item Search

Two different methods were proposed. One is traditional text-based and another is novel bio-inspired method.

1.1 New bio-inspired method

*This work was supported by National Natural Science Foundation of China under Projects 61101212 and 90920001, and by Fundamental Research Funds for the Central Universities, and Network System and Network Culture Foundation of Beijing.

Inspired by human attention, recognition and binding mechanisms, we proposed a novel approach to KIS task in TRECVID 2011. In this approach, a query topic is first parsed by a text analyzer to produce several search cues, and then the cue-based bottom-up saliency map and the top-down cue-guided concept/object detection are fused and refined by the aid of context cues. With this new bio-inspired method we achieved better results than those obtained by the high-level feature/concept-based method in TRECVID 2010.

A The proposed framework

The proposed KIS framework is shown in Figure 1. It mainly includes five parts: a bottom-up attention model for determining salient regions, a knowledge base containing various pre-trained object/concept (such as person, car) detectors, a SOM (Self-Organizing Maps) network to map known-item keywords into seven image-related classes, a SVM scene classifier for data filtering, and a fusion module to perform content-based retrieval, results fusion and ranking.

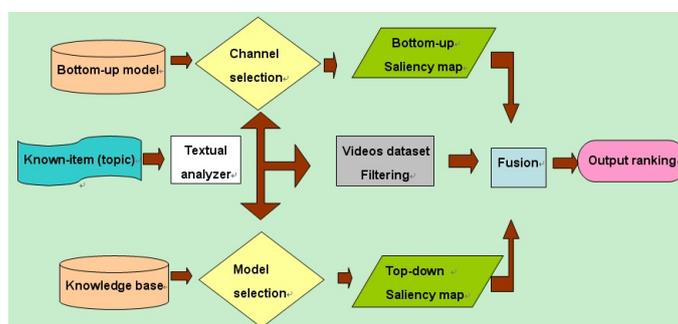


Figure 1.1. The KIS framework of bio-inspired approach.

B The SOM Network of Words

The human brain can easily parse text sentences and map the semantic information into visual images. Therefore, if relationships between keywords of search topics and image features/concept can be established, the semantic gap could be narrowed.

In our system, a textual feature vector including noun similarity, noun hierarchy, adjective similarity etc. is extracted to describe each keyword. A SOM network is then trained to cluster keywords into seven categories, including color, texture, shape, person, vehicle, position relations between objects and specific semantic words. The first three categories will give the bottom-up saliency cue, next two give the concept detection (top-down) cue, and the last two respectively give the context and data filtering cues. During KIS search, the query keywords are automatically classified by the trained SOM.

C The Attention Model

According to the two-channel (Where-What) theory of human visual system, in our framework, a bottom-up combined with top-down attention model is built up to detect informative objects described in the search topic. Firstly, the corresponding channels (color, shape and spectral residual) of the attention model are weighted with the bottom-up cues to locate potential salient regions. And then, proper models are selected from the knowledge base according to the top-down cues to locate objects/concepts interested. Finally, Bayesian inference is employed to fuse the bottom-up and top-down saliency maps, and the context cues are used to refine the results.

D Scene Classification

In order to enhance the search speed and performance, a scene classifier based on Gist and SVM is employed to classify video scenes into two categories: outdoor and indoor. In addition, a black and white video detector is also developed. Both classifiers are used to filter out irrelevant videos.

1.2 Text-based method

The proposed automatic text-based search system is consisted of several main components, including text pre-processing, keywords extracting and processing, text-based retrieval, results fusion and re-ranking. The framework is shown in Figure 1.2.

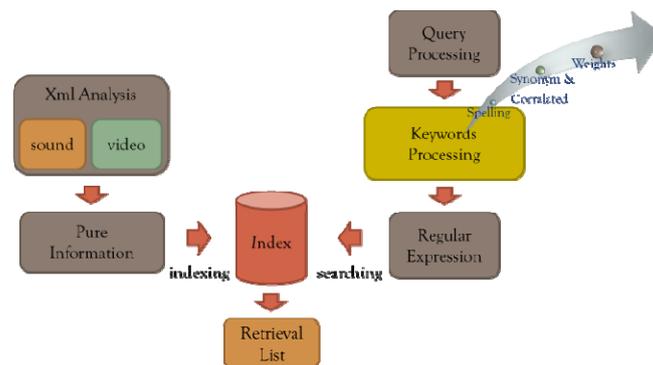


Figure 1.2. The framework of text-based approach.

A Text-based Ontology Construction

A text-based ontology is constructed manually. Using this tree-like lexicon, we can obtain information on whether words are particular colors, language, places, specific terms, sound etc. in search topics.

B Text Processing

- Spelling mistakes in queries as well as in metadata and ASR data were corrected.
- With the aid of NLP tools, we eliminated much redundant information in topics and metadata, and extracted the keywords according to words' weights.
- Using Youdao dictionary, we expanded the extracted keywords by finding their synonyms and correlated categories.
- Finally, we recomputed the weights of each processed-keyword based on their importance, and then changed them into regular expressions to search.

C Text-based Retrieval

In text-based retrieval, keywords from sound data and metadata were first converted to a stream of plain-text tokens, as explained in Figure 1.2, and then were sent to Lucene to build text indexings, and the same process was applied to search topics. Finally, the sound data and metadata were searched individually and search results were combined and re-ranked.

1.3 Results and analysis

Figure 1.3 shows final KIS results of all automatic search runs, and two brown bars are performances of our two different runs: the higher one is purely text-based search and another is content-based retrieval proposed above.

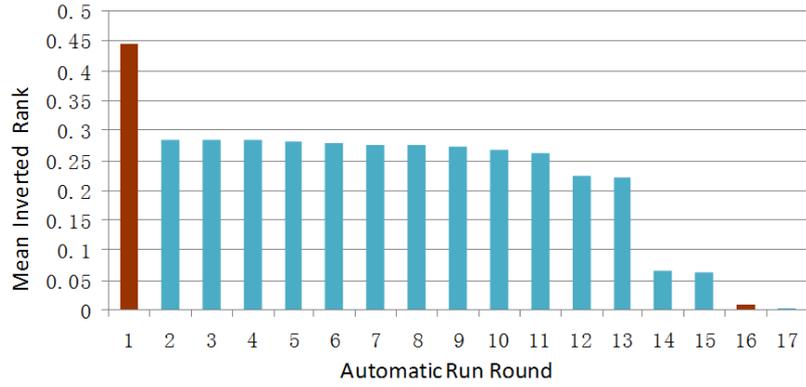


Figure.1.3. Results of all automatic search runs.

From this figure, we can see that our text-based approach achieved the best result of 45.5% MIR among all automatic search runs, and the proposed bio-inspired method obtained 0.9%. However, it should be pointed out that among all 391 search topics, only 53 topics are searched with the proposed method, and remaining results are randomly generated. If only considering the 53 topics, 6.56% MIR would be obtained, and this result is shown in Figure 1.4. Furthermore, we find that the performance is obviously enhanced from 0.4% to 0.9% when comparing with concept-based approaches of our team and MCG team last year. Therefore, we believe that the proposed bio-inspired method is promising if the attention model and knowledge base are further improved.

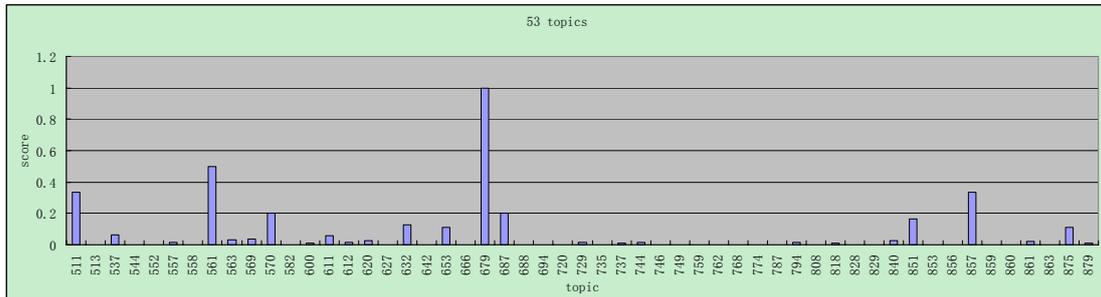


Figure 1.4. Results of 53 search topics.

2 Instance Search

The proposed automatic instance search system is consisted of several main components, including visual query pre-processing, keyframes and features extraction, keyframes retrieval, multimodal fusion and results re-ranking. The framework of our INS system is shown in Figure 2.1.

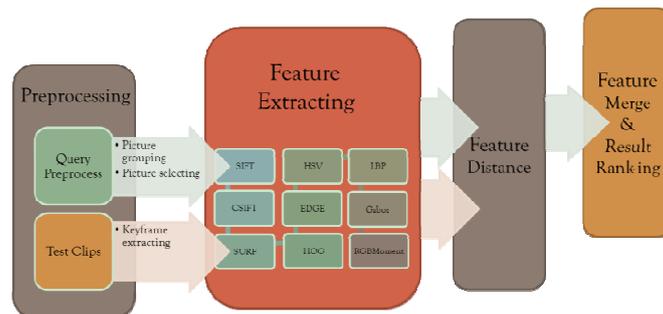


Figure 2.1. The framework of INS system.

2.1 Feature Selection

Since no unique visual feature can represent all information contained in a keyframe, and no given visual feature is effective for all topics, we extracted several visual features at regional and global levels[1, 2], the details of which are listed in Table 2.1.

Table 2.1: Selected visual features

Features	Description
HSV Histogram	HSV color histogram for global partition
RGB_Moment	225 dims RGB color moment feature for global partition
SIFT	SIFT feature and BoW method with 1000 visual words
SURF	SURF feature and BoW method with 1000 visual words
CSIFT	CSIFT feature and BoW method with 1000 visual words
Gabor Wavelet	3-scale and 6-direction Gabor feature with 3*3 regional partition
EDH	145 dims histogram by concatenating global and regional EDH
LBP	256 dims histogram of each LBP code with global partition
HOG	2520dims histogram with 10*7 regional partition and 9 directions

2.2 Keyframes Clustering for Shot Category

Shot keyframes were extracted to cluster and classify into two different categories: long shots and close shots. For example, in Figure 2.2, two viewpoints are shot in the same scene, but close shots are usually more important for INS. Therefore, during the results fusion of INS, we assigned a higher weight to these situations.

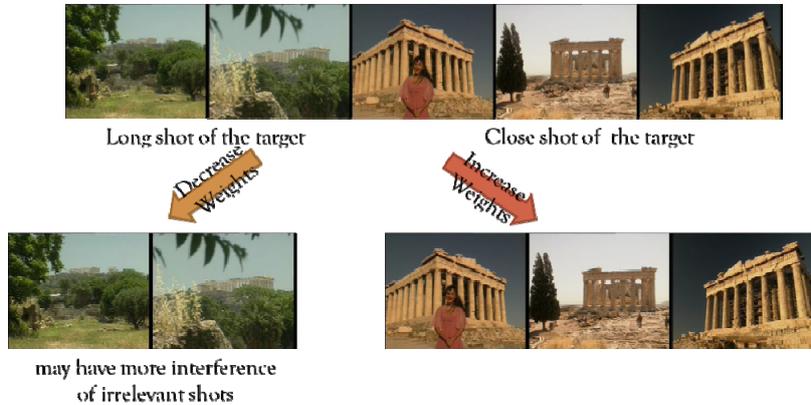


Figure 2.2. Picture grouping.

2.3 Similarity Computation

We used various normalizations for visual features and employed different distances to compare feature's similarity. For HSV, RGB_Moment, Gabor, LBP, Edge and HOG, we used sum-normalization, while SIFT, SURF and CSIFT Gaussian-normalization was used. Furthermore, we used Bhattacharyya distance instead of Euclidean distance to compute HSV similarity.

2.4 Results and Analysis

Figure 2.3 shows final search results with 3 runs of different feature merging strategies. One run achieves 0.407 MAP which indicates that our method is feasible.

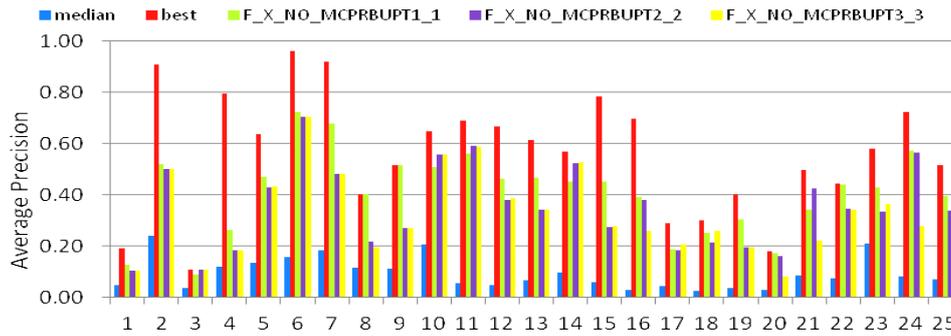


Figure 2.3. Our results VS median and the best.

3 Semantic Indexing

This year, we submitted 3 runs of semantic indexing task. Compared with the system of TRECVID 2010, some changes were made:

- Feature: In threavid 2011, four SIFT-like features, SIFT, CSIFT, OpponentSIFT and RGBSIFT, are adopted.
- Kernel function: During the SVM training, we applied Chi-square kernel [1] rather than RBF Kernel.
- Word Projection: We adopted soft assignment of visual word projection instead of hard assignment to incorporate ambiguity in visual word representation.

3.1 System Framework

Our visual based semantic indexing system consists of three components: feature extraction, classification and fusion, which is shown in Figure 3.1.

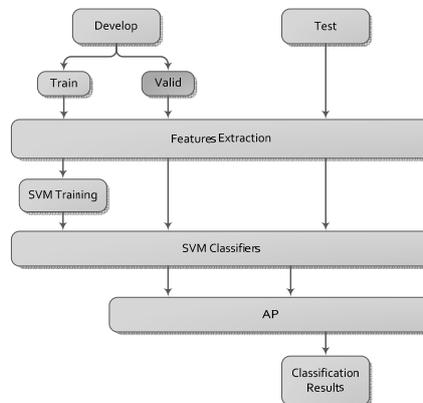


Fig 3.1. The framework of semantic indexing system

The frames labeled by active participants were divided into training set and validation set, the former was used to train models and the latter was for evaluations. First of all, we trained a series of models on the training set for each concept, and then average precision (AP) was generated on the validation set, which was an evaluation criterion for models trained before and linear weighted parameter in the fusion step later. The testing set was processed as before and the final SIN result was generated with three fusion methods.

3.2 Features Extraction

4 different SIFT-like visual features, which are listed in Table 3.1, were extracted to describe the video

contents. Then a soft projection representation [3] was used to generate our codebook model.

Table 3.1 Selected low-level features

Features	Description
SIFT	The SIFT feature describes the local shape of a region using edge orientation histograms
CSIFT	The SIFT feature calculated in the color invariance space
RGBSIFT	For the RGB-SIFT, the SIFT feature is computed for each RGB channel independently
OpponentSIFT	OpponentSIFT describes all the channels in the opponent color space using SIFT features

3.3 Multimode Fusion

For each concept, three fusion strategies were employed, one with all features, one with best three features, and the other one with best feature. For each concept, the final retrieve list was based on the probability that were generated by SVM classifiers and linear weighted with APs.

4 Copy Detection

This year, we focused on video copy detection, and audio copy detection was implemented on the basis of the framework in last year.

4.1 Video Copy Detection Based On SIFT

The framework of our system is shown in Figure 4.1. It mainly includes two parts, one is the online processing part and the other is the offline processing part. The former was used to query videos, and the latter to reference videos.

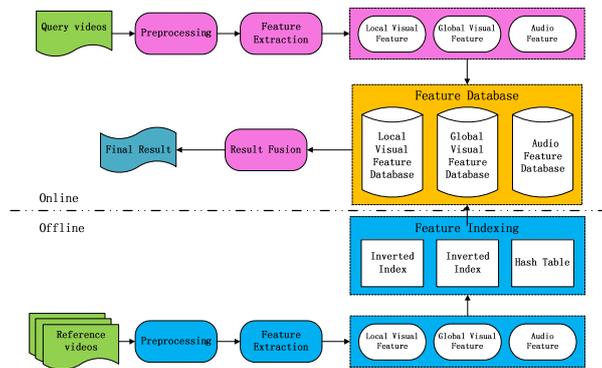


Figure 4.1. Framework of video copy detection system.

4.2 Detection for PIP and Feature Extraction

Since original video and audio information cannot be used to find copies directly. Robust video features and audio features should be selected to construct feature databases for copy detection. In addition, PIP seems to interfere with robust visual feature extraction, so in this section a feasible method is introduced for PIP detection.

A PIP Detection

For PIP transformation, the resolution of the original video is too small that common methods cannot perform well. Consequently, we detect this transformation separately. Figure 4.2 shows the flow chart of our method. Firstly, uniform sampling, 1 frame out of 25 frames, is used to obtain the key frames. Then Canny operator is applied to get the edge graph of each key frame and each of them are aggregated into a whole map. Later Probabilistic Hough Line Detection algorithm [15] is used to get potential line segments. Finally, rule-based method is proposed to get the rectangular regions of PIP.

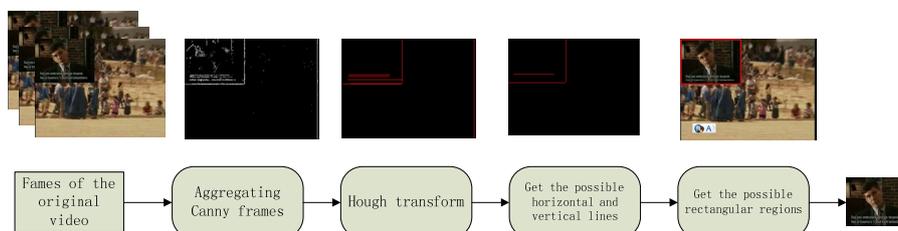


Figure 4.2. PIP detection algorithm.

B Global Feature Extraction

Color correlograms and LBP are extracted from each video frame. These two global features indicate the color and texture feature of the frame separately. These global visual features can be calculated efficiently and they are robust to some transformations.

4.3 Rank-based Fusion Scheme

As each query can only have one reference in the database, the results based on different features must be merged to get the most possible reference. The rank-based scheme is shown in Figure 4.3. First, we get at top 5 candidate videos from each feature searching scheme. Then the intersections of the results are assumed the accurate candidates which are subtracted from original candidates. Then the remaining candidates are scored by a scaled rule of their original feature score and the candidate which receives highest score is added to the final results.

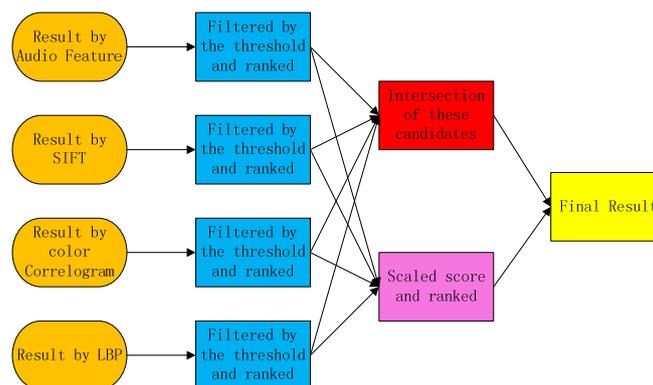


Figure 4.3. Rank-based fusion scheme.

4.4 Results and Feature Work

From the CBCD experiments result of TRECVID 2011, we found that:

- F1 is better than the median, but NDCR is high.
- Detection for PIP is effective.
- Audio copy detection is not effective.

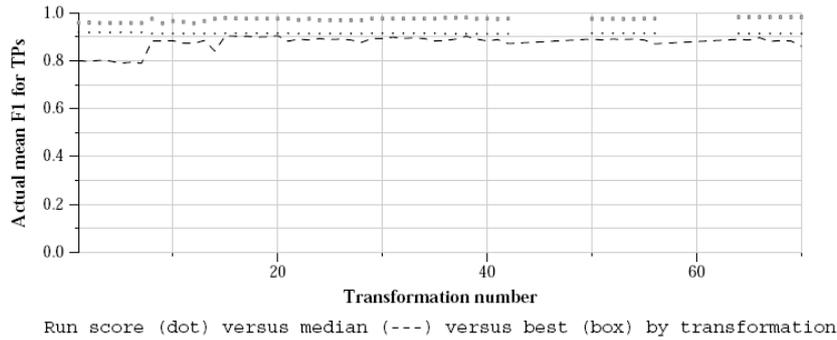


Figure 4.4. CBCD results.

5 Surveillance Event Detection

This year, we mainly focused on the events of PersonRuns, PeopleMeet, PeopleSplitUp, ObjectPut, Embrace and Pointing. Our system adopted different algorithms in detecting these events accordingly.

5.1 PersonRuns Detection

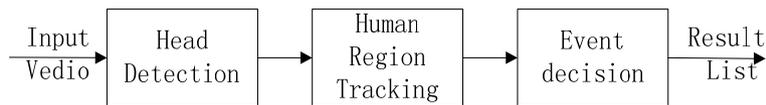


Figure 5.1. The diagram of human action detection.

Our system for PersonRuns Detection is typically the same as, except several modifications, that of the last year, which consists of the following parts: human head detection, human head region and whole body region tracing, trajectory analysis and PersonRuns decision. The main steps are as follows:

Step 1: Human head detection. We firstly find the possible points of head-top according to the gradient of both the video frames and the foreground image generated by it. If a pixel has an almost vertical gradient in either video frame or foreground image or both, it will be marked as a possible point of head-top. After that a region of interesting (ROI) is obtained from each point then the HSV feature of the head region as well as the whole body and histogram of gradient (HOG) feature are extracted. Finally several SVM classifiers are employed to decide whether or not the ROI is a human head. Hence an object list of head region could be generated by arranging each of the detected human heads. Each object is described by its head region HSV feature, whole body HSV feature and HOG feature.

Step 2: Human head tracking. In the subsequent frames, the system detects new human head region and compares the feature of these regions with that of the objects in the list. For the matched object, system replaces its features by the new one detected from current frame. While for the ones that mismatched, the system searches the adjacent region around the prediction position to find the corresponding object. In that way, the system is able to trace the object from one frame to another and gain the trajectory of objects.

Step 3: Trajectory analysis. The information of speed, distance, acceleration and linearity can be obtained by trajectory analysis. The decision score can be calculated by fusing these informations.

Step 5: PersonRuns decision. The scores get from Step 3 can decide whether an event of person runs occurs.

5.2 PeopleMeet and PeopleSplitUp

PeopleMeet and PeopleSplitUp detection is based on the low-level features. In such cluster scenes, using tracking based method to detect multi-persons activities is useless. Thus, we applied MoSIFT[4] features as our low-level features which presents the large video data in an elegant way. After extracting the features, bag of the words method is used to find the meaningful features centers to avoid divergence. In consideration of combing these structural features, we use the SVM-HMM[5] to train and classify the samples generated by the sliding window method. It is notable that assemble classifiers is applied to vote the final result. The whole process is described as Figure 2

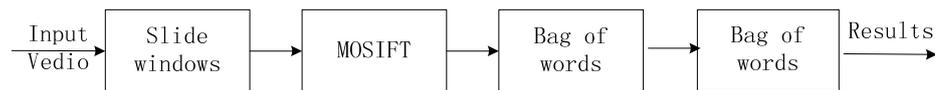


Figure 5.2. The diagram of PeopleMeet and PeopleSplitUp detection.

5.3 ObjectPut, Embrace and Pointing Detection

For ObjectPut, Embrace and Pointing events detection, the frame work is described in figure 3. Our approach for recognizing these three events mainly consists of three stages. For each input video, firstly, we limit our attention on regions contain movement which simply utilize frame difference method. We implement Bruhn's real-time dense variational optical flow algorithm to compute optical flow of the attention region. Then for each pixel belongs to the interesting region, we compute several features derived from the optical flow field and adopt log-covariance matrix as our feature descriptor. Finally, support vector machine (SVM) classifier with radial-basis kernel is adopted for events classification.

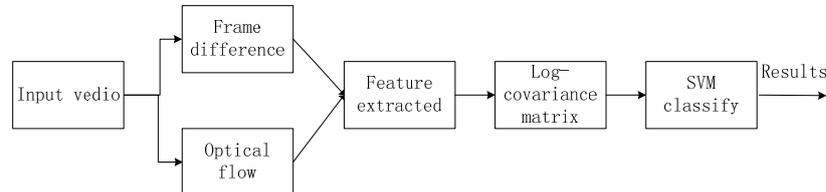


Figure 5.3. the diagram of event detection for ObjectPut, Embrace and Pointing events.

5.4 Conclusions

There are some problems in our SED algorithm needed to be solved. For individual behaviors, respective action should be considered which will be improve accuracy to these event detections in the future, for example, ObjectPut, Pointing and PersonRuns etc. As the single person detection and tracking are the two most important parts in the event diction system, so human head would be more accurately detected, and also more robust tracking method should be considered. The second one is to research more discriminative features to describe human motion. The third is that we need to decrease to the rate of false alarms which influence the system performance deeply

References

- [1] Xiaoming Nan, Zhicheng Zhao, Anni Cai et al, "A Novel Framework for Semantic-based Video Retrieval", ICIS 2009.
- [2] Zhicheng Zhao, Yanyun Zhao, Zan Gao, Xiaoming Nan et al, "BUPT-MCPRL at TRECVID 2009", In: Proceedings of TRECVID 2009 Workshop.

- [3] C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.
- [4] M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Computer Science Department, 2009.
- [5] Thorsten Joachims. SVMHMM tool package available at http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html.